The Influence of Education Levels on Income Inequality

Paula Cuadrado

Olivia Fulmore

Elizabeth Phillips

Abstract

This paper explores the role of education in income inequality. The data comes from the United Nations Development Programme Human Development Reports and the Organization for Economic Cooperation and Development. We started with a simple linear regression, using the sample countries' Gini index as a proxy for income inequality and the average years of schooling as a gauge of education levels. The Gini index is regressed on the average years of schooling to determine if low education levels lead to income inequality. Many other factors may contribute to income inequality, so additional models perform multilinear regressions. The dependent variable, Gini, stays the same, but we add new independent variables such as median age, percentage of population engaged in vulnerable employment, log of GDP per capita, government education expenditure, government health expenditure, foreign direct investment inflows, and a dummy variable for OECD countries. The purpose is to inform public policy on the allocation of educational resources in countries seeking to combat inequality.

I. Introduction

This paper will explore the possible relationship between a country's level of education and its income inequality. The purpose is to test how additional years of schooling for citizens impact a country's income inequality. If there is a strong, negative relationship found between average years of schooling and the Gini Index, then it may be wise for countries with low education levels to consider redistributing resources toward groups of citizens that are currently receiving few years of education. Otherwise, a country seeking to solve its income distribution issues ought to choose another policy prescription.

Hypothesis

Income inequality can have damaging economic effects, and when the inequality is persistent over time, social mobility is threatened (OECD, 2019). Social mobility indicates how possible it is for an individual to move from one income bracket to another within a country. When mobility is low, a country forfeits the economic benefits of fully realizing the potential of individuals at the bottom of the income distribution. Disadvantages are passed on from parents to their children, creating cycles of economic stagnation for portions of the citizenry. This lowers overall economic efficiency. Meanwhile, individuals at the top of the income spectrum will remain there, allowing them to benefit from the unequal system as well as incentivizing them to maintain the status quo. Thus, as the wealth concentrated at the top of the income spectrum swells, the balance of economic power swings in favor of perpetuating the inequality. This means that it is critical that income inequality is addressed sooner rather than later, while there is still enough support available to throw behind the changes. So, it is worthwhile to investigate exactly which policy changes ought to be made in order to achieve this outcome. Education is one of many factors that might influence social mobility; other possible factors could include government welfare programs, foreign investment, or the existing level of economic prosperity.

According to the United Nations Sustainable Development Goals, countries should strive to both reduce inequality and increase the availability of education. Annan-Diab (2017) explains the emphasis that world leaders are placing on these factors. We expect to find a negative relationship between average years of education and income inequality at the country level; meaning, as citizens' educational attainment goes up, countries become more equal on the basis of income.

II. Literature Review

Park (1996) starts out by examining the inverse-U structure of the Kuznets Curve and bringing about a new interpretation of the curve. In this case, there will be more weight on education variables, particularly focusing on level of education and income. Throughout the study conducted in this article, Park incorporates the Gini Index as well as income as a measure of the dependent variable run by its models. The education variables used to explain income were separated into four different categories: enrollments at different levels, mean/median years of schooling, rate of return at the different levels, and dispersion of educational attainment. An interesting finding from this paper was the negative effect education inequality and level of schooling have on income distribution, when used in conjunction, as explanatory variables. In order for the regression to show a positive effect between level of schooling and income distribution, the education inequality variable must be removed. The reasoning given for this phenomenon is the high correlation that is present between the level of schooling and the per capita income. Along with a high correlation, there is collinearity between the level of schooling and education inequality. The reason for this collinearity is due to the educational inequality variable already containing level of schooling within it.

Judson (1998) examines the response of economic growth to the production of human capital through education. Primarily, this paper is concerned with the allocation of educational resources. Judson makes multiple assumptions before constructing his model. It is stated as fact that years of education yield diminishing marginal returns; thus, investment in primary school has a larger economic return than investment in higher education does. However, this fact does not necessarily hold when returns from secondary education are compared to those from higher education. This claim is informative for our own research. Another interesting technique used in Judson's paper was the allowance for "revelation of talent". Individuals are not all equally talented, so the more talented ones should receive more education as they reveal themselves to be worthwhile investments. This strategy would defeat the aims of our research, as our goal is to reduce inequality, rather than to maximize absolute growth. By including this dimension, Judson creates a model that can be used to determine if a country's allocation of educational resources is efficient or not. After determining efficiency, Judson assesses the relevance of a country's efficiency score.

Sylwester (2002) starts by pointing out an assumption that has been very often overlooked when creating policy to combat education inequality: children from low income families are actually attending the schools governments are funding. The main concern highlighted in the article is that although there has been an increased resource allocation towards education, many countries have not seen a difference in their existing, unequal income distribution. In this study, Sylwester divides the countries into OECD and

non-OECD and makes a comparison of the results gathered from both. He finds that countries that are part of OECD groups did experience an equalizing effect on their income distribution based on an increase in resource allocation towards education. The same effect was not found in the countries outside of this group. An explanation given for this difference in income convergence stands in the initial overlooked assumption; children who come from lower income families may have to work in order to support their families and therefore cannot afford to attend school. For many years, it was taken as given in the research of development economists that increasing access to education will certainly reduce inequality. However, more recent literature has shown that this may not hold true in countries with low development from the start. This is because the opportunity cost of attending school is too high, meaning that poor students' tax dollars are sometimes spent on public school while the students themselves cannot afford to take advantage. When this is the case, income inequality actually worsens. We would like to build on this finding by determining if inequality could be decreased if children received more years of education across income brackets.

A population's wealth may be related to its age according to the St Louis Fed (Vandenbroucke 2017). Researchers examined the connection between these two variables and made projections about what the effect would be on wealth inequality. Based on the studies, they found that in the United States the age group holding the largest percentage of wealth is individuals ages 55-64 years, even though they only comprise 16% of the total population. Individuals 65 years and older currently account for only 15% of the total population, but they are projected to make up 30% of the population by 2030. Income accumulation and retirement savings are attributed to the differences in wealth per age group and are used to further explain why the oldest members of the population, despite making up a smaller overall percentage, hold the largest percentage of wealth. This study makes a clear case for increasing wealth with age, which we will keep in mind when considering factors other than education that may be important to explaining inequality.

Abdullah (2015) conducts an empirical analysis on whether education reduces income inequality. They start out by examining social and political issues inequality poses upon a society as well as its effects on growth and overall development. In terms of education they examine the rates of return to education as well as the effect of education on income shares and income inequality. They examine the labor force and the changes the growth in the supply of educated workers has on it. Their claim is that a larger percentage of educated workers in the workforce in the long run is expected to reduce income inequality. The reason this would lower income inequality is because with an increased supply of skilled workers the wage premium would become lower, which would buy default lower income inequality. Another factor they touched on relating to inequality in education is the level of investment in education

as well as government intervention. While education subsidies were created to give opportunities for poor children to access education, the effectiveness of the subsidies is not entirely clear. In order for public spending in education to reduce the income gap there must be equal access to education. The perceived reason for its ineffectiveness is due to the fact that an increase in education spending may not benefit lower income communities if they are unable to attend school because they cannot afford to do so.

Cingano (2014) analyzes a study regarding trends in income inequality and their impact on economic growth. Among OECD countries, the gap between the rich and the poor is continuously increasing. There has been a rise in income inequality which benefits those at the top during prosperous years to the detriment of those at the bottom. Based on the data of OECD countries over the past 30 years, the models show evidence of income inequality negatively affecting growth. The authors assert that when applying policies designed to reduce income inequality the focus should not be to simply improve social outcomes but also foster long-term economic growth. The most important target of these policies is promoting equality of opportunity to increase educational attainment as well as quality of education. We received two key takeaways from this article: there may be an important distinction to be made between developed and developing countries, and length of education does not speak to its quality. We plan to incorporate variables in our regressions that will account for this.

Annan-Diab (2017) Talks about the need of improving education for sustainability in order to meet the Sustainable Development Goals set in place by the UN in September 2015. The SDGs seek to end poverty, protect the planet, and allow for a more prosperous society. Education has an entire goal dedicated to its continuous improvement. Needless to say, education is part of sustainable development; this paper acknowledges the importance of an interdisciplinary approach to tackling education inequality. Education inequality is captured under two different SDGs, reduced inequalities and receiving a quality education. The key to achieving these goals is by taking advantage of interdisciplinarity and creating well rounded curriculum to help people succeed in any sector they wish to advance in. We found consideration of the SDGs to be highly probative to our own research; these are the goals of modern governments, and it is important for our research to be in touch with the work that global leaders are engaged in.

Aaberge (2014) states that when studying income inequality, it is more relevant to look at the distribution of lifetime income rather than the distribution of current income. These findings contradict a majority of empirical studies that relate to income inequality, as those are based on income of one or a limited number of years. This paper states that taking a closer look at career-long incomes will allow the minimize effects from “lifecycle bias”. They believe standard of living depends on lifetime income; individuals can borrow from year to year to even out life cycle changes. The reason lifetime data is not often used when conducting studies is because it is difficult to collect, and thus not easily found. The

lifecycle bias is made up of two components: income mobility and heterogenous age income profiles. This study found that "income mobility reduces lifetime income inequality by 25%" while heterogenous age-income profiles contribute positively to income inequality when measured later in the work life cycle and vice versa. When studying intergenerational income mobility, regression models that use current income as opposed to lifetime income will produce results that are inconsistent. Because there is no way of knowing at what age the current income used in the regression is coming from, it will fall into the life cycle bias trap.

Our research takes note of the work mentioned in Sylwester's paper by adding in OECD and Non-OECD countries as a dummy variable in our multiple linear regression models. We were not the only ones to pay attention to differences between OECD and non-OECD countries as this was mentioned in Cingano's study of income inequality and economic growth. Allowing two categories of development within countries to be measured separately helps eliminate bias within our models and thus allows us to achieve more robust results. We also responded to Cingano by including a variable for government expenditure on education as a proxy for quality.

III. Data

Income inequality can be affected by many different variables. In order to capture the full story of income inequality we decided to include a diverse set of variables in order to reach more dimensions of a country. We analyzed 150 different countries in order to expand our random sample and include countries across the equality and income spectrum such as Niger and the United States. We felt as though this would provide a more holistic and accurate view of income inequality. The chosen variables had readily available data from years spanning between 2010 to 2017.

Gini Index: The Gini Index serves as our dependent variable. The Gini Index is a measure of income distribution and more specifically income inequality. The index has two extremes of 0% and 100% , where 0% represents perfect equality, and 100% represents perfect inequality. We utilized the United Nations Development Programme Reports (UNDP) in order to obtain data on this index for all 150 countries between the years 2010-2017.

Average Educational Attainment: Many countries require child education or an equivalent to it. We decided to use average educational attainment, meaning average years of schooling, as our main independent variable so we could address our original hypothesis, which broadly questioned the impact of education on income. In terms of this project, we wanted to analyze the response to average educational

attainment increases from income inequality. With data from the UNDP Reports, we analyzed the average years of schooling in 2017. The UNDP measures average years of schooling through the average number of years of education received by people ages 25 and older, converted from education attainment levels using official durations of each level.

Median Age: Median age is the second independent variable. Utilizing the literature from the Saint Louis Federal Reserve, we anticipated that there would be a strong correlation between median age and Gini Index. We wanted to see how trends found in the United States would translate to a global scale and see if lower median age countries such as Kenya would have greater income inequality than higher median age countries such as Japan.

Government Expenditure on Education: In addition, we utilized government expenditure on education to measure how much a country spends on education. This was an interesting statistic because it serves as a proxy for how much the country values education and invests in it. It could also be argued as a proxy for quality. Additionally, government expenditure on education is correlated to average educational attainment; however, average educational attainment does not fully explain how much the government invests in education that is why we included it in order to get rid of omitted variable bias.

Vulnerable Employment: Vulnerable employment is the percentage of people employed within a country that are considered unpaid workers, such as home makers, or are self-employed. This variable captures people that do not have a stable source of income, therefore affecting the income inequality within a country. If a country has high vulnerable employment, then it may be more likely to have higher income inequality due to disruptive factors such as uncertainty.

Health Expenditure: Health expenditure measures the percentage of GDP a country spends on health related services or goods such as research and development of new medicines. This variable excludes capital health expenditures such as buildings and machinery. We decided to include health expenditure because health equality may coincide with income inequality. When income inequality is relatively high, we expect the percentage of GDP a country spends on healthcare to be low.

OECD: The Organization for Economic Cooperation and Development, OECD, was founded in 1961 as a promoter of progress, world trade, and thoughtful economic decision making . There are currently 36 member countries, and membership may be considered a proxy for economic development. We made OECD membership our dummy variable because we expected a correlation with income inequality in this country.

GDP per Capita: In order to capture a country's economic output relative to population, we used GDP per capita as an independent variable. GDP per capita measures the productivity of a country by dividing its gross domestic product by the total number of people living within this certain country. This serves a proxy for a country's standard of living, which would ultimately affect its income inequality. We went a step further and took the natural log of GDP per capita. GDP varies dramatically from country to country, and if left alone, its inclusion could increase heteroskedasticity. In order to conform with Gauss Markov Assumption 5, we took the log of GDP per capita, which made it more statistically significant.

FDI inflows: FDI inflows stands for federal direct investment inflows into a certain country. This variable measures the reinvestment of earnings, sum of equity capital, and other capital both long term and short term that are being invested in a particular country. More specifically the variable is expressed as a percentage of GDP in order to easily compare across countries because FDI varies drastically like GDP per capita. We decided to include this variable because the more FDI a country receives, the more equal the country is thought to be because of the resources that become available to the country. In addition, to avoid omitted variable bias, we added FDI inflows in order to capture the full story around FDI in a country.

Data Sources: The United Nations Development Programme Human Development Reports (UNDP) will serve as the source for our data as it has multiple indicators on inequality around the world, including education, government expenditure, and income inequality. The UNDP Reports utilize many different surveys around the world in order to gather the data within different countries. These surveys employ random samples of the populations at hand therefore contributing and complying to Gauss Markov Assumption two of random sampling. We decided to use the UNDP Reports because of their mission to move toward sustainable goals and diminish inequality. Finally, we utilized the Organisation for Economic Cooperation and Development for our dummy variable in order to divide our countries into developed vs undeveloped. The specific sources used within each report and variable are listed in the appendix.

Descriptive Statistics:

The summary statistics listed in Table 1 give insight to the bigger picture of each of our variables. Within these summary statistics we observe the observation number, mean, standard deviation, minimum and maximum for each variable. These summary statistics provide a prospective for a certain country in regards to the rest of the world.

Gini Index: The Gini Index serves as our dependent variable. The mean Gini Index is 38.17 with a standard deviation of 8.13, meaning that income inequality varies within countries, but is relatively low. We know the standard deviation is low because the coefficient of variation is 0.213, which is less than one, therefore providing a benchmark. Further, the minimum Gini Index is 16.6 (Azerbaijan). This implies that Azerbaijan has low income inequality. On the other hand, the maximum Gini Index is 63 (South Africa). This implies that South Africa has high income inequality. For reference, the United States Gini Index is 41.5, which is actually higher than the average, meaning that our country has higher than average income inequality.

Average Educational Attainment: 187 observations were taken for Average Educational Attainment with a mean of 8.52 years of schooling and 3.10 standard deviation. Once again, the coefficient of variation is relatively low, implying that there is not much variation within the data. The minimum is 1.5 years (Burkina Faso) and the maximum is 14.1 years (Germany). Overall these statistics reflect the vast differences in value of education between countries.

Median Age: In 184 observations, the median age was calculated with a mean of 28.79 and standard deviation of 8.86. The coefficient of variation is relatively low providing the inference that there is not much variation within the data provided. Further, the minimum median age is 14.9(Niger), while the maximum median age is 46.3 (Japan). This builds the story of populations growing old versus young populations. This also might correlate to the education and opportunities that the people within these countries have. For example, in Niger, the population is relatively young, perhaps due to difficulties in access to birth control or a lack of education for women. All of these factors can ultimately affect income inequality.

Government Expenditure on Education: South Sudan has the lowest government expenditure on education with a value of 1.8% of GDP, while the Federated States of Micronesia has the highest government expenditure on education with a value of 12.5% of GDP. Out of 139 observations, the mean is 4.69% with a standard deviation of 1.69%. It would appear that governments invest relatively little in education on average, which may be contributing to income inequality.

Vulnerable Employment: 92.4% of Burundi's population is either self employed or works without pay. Other nations at the top of the list include Chad, Niger, and Sierra Leone. All of these countries are found in West Africa, which is known for having high income inequality. We see a correlation between vulnerable employment and income inequality. On the other hand, only 0.2% of Qatar's population is self

employed or works without pay. Further, approximately 36.95% of people globally work for themselves on average with a standard deviation of 26.29%.

Health Expenditure: 2.5% of South Sudan's GDP is spent on health expenditure, while 18.3% of Sierra Leone's GDP is spent on current health expenditure. The mean of current health expenditure around the world is 6.7 with a standard deviation of 2.74. This creates the narrative of how countries spend relatively low amount of money of health expenditures, which can contribute to income inequality.

OECD: OECD serves as a dummy variable in order to separate the countries into developed vs undeveloped countries. In the summary statistics we observe 189 countries where most countries are not apart of the OECD with a mean of 0.179 therefore we are considering most of the countries that we used undeveloped. The minimum and maximum of this variable are 0 and 1 respectively because of its purpose of being a dummy variable.

GDP per capita: We took the natural log of GDP per capita in order to reduce heteroskedasticity. With this being said within our 189 observations, the mean was 9.24 with a standard deviation of 1.18. The minimum is 6.49 (Central African Republic) and maximum is 11.67 (Qatar).

FDI inflows: Finally, we measure FDI inflows as a percentage of GDP. Over 187 observations, we calculated the mean of FDI inflows into a country was 4.45% with a standard deviation of 7.77%. With this being said some FDI inflows actually came up negative with Iceland being the minimum at -29.4 and Cyprus being the max at 48.6. With this all being said, FDI inflows' coefficient of variation is above 1, indicating wide variation and implying that there is high variances between countries.

Summary Table of Variables

The table below shows the summary statistics for each variable within our regressions.

Table 1: Summary Statistics

Variable	Observations	Mean	Std. Dev.	Min.	Max
Gini	154	38.17	8.13	16.6	63
avgschool	187	8.52	3.10	1.5	14.1
medage	184	28.79	8.86	14.9	46.3
vulemp	179	36.95	26.29	0.2	92.4

loggdppercap	185	9.24	1.18	6.49	11.67
govtexpendeduc	139	4.69	1.69	1.8	12.5
newhealthexpend	184	6.74	2.74	2.5	18.3
OECD	189	0.179	.385	0	1
FDIinflows	187	4.45	7.77	-29.4	48.6

Gauss Markov Assumptions

1. MLR.1 Linear in Parameters: Both Simple and Linear Regressions that are run in Stata are linear in parameters. We checked that they are linear in parameters from our equations where no variables are multiples of others. Therefore, our equations are:

   SLR: $\widehat{GINI} = \hat{\beta}_0 + \hat{\beta}_1$ avgschool$+\hat{\mu}$

   MLR: $\widehat{GINI} = \hat{\beta}_0 + \delta_0$ OECD$+ \hat{\beta}_1$ avgschool$+ \hat{\beta}_2$ medage$+ \hat{\beta}_3$ vulemp$+ \hat{\beta}_4$ loggdppercap$+ \hat{\beta}_5$ govtexpendeduc$+ \hat{\beta}_6$ newhealthexpend$+ \hat{\beta}_7$ FDIinflows$+\hat{\mu}$

   As seen above, the equations are linear in parameters therefore they meet the first Gauss Markov assumption.

2. MLR.2 Random Sampling from Population: We surveyed 150 countries from the UNDP Reports. We picked the maximum number of countries we could use in order to ensure random sampling from both high and low income countries. Therefore our sample that we used for our statistical analysis is random and meets the second Gauss Markov assumption.
3. MLR.3 No Perfect Collinearity: The variables we chose have correlation to one each other as seen in the correlation chart. Some variables such as Average Educational Attainment and Median Age have high correlation to each other, but are not perfectly correlated meaning as one increases by one unit, the other one exactly increases by one unit as well. Since none of our variables have a perfect one to one correlation, our regressions meet the third Gauss Markov assumption of no perfect collinearity. *See Appendix: Output 1 for correlation coefficients.*
4. MLR. 4 Zero Conditional Mean: First, Zero Conditional Mean checks whether the error term μ has an expected value of zero given the values of the independent variables. This means that the

meaning of the error term μ does not depend on any of the independent variables. This ultimately checks for any omitted variables within the model that would affect the dependent variable. After extensive research, we have concluded that μ is zero and there is no omission of any variables concluding that our models meet the fourth Gauss Markov assumption of zero conditional mean.

5. MLR.5 Homoskedasticity: This assumption assumes that the variance for the error term μ is similar all independent variables. This is shown by scatter plots of the independent variable in regards to the dependent variable. *See Appendix: Figure 1 for scatter plot and trend line to illustrate the variances.*

IV. Results

Model 1: Simple Regression Model

Equation:

$$\widehat{GINI} = \hat{\beta}_0 + \hat{\beta}_1 \text{ avgschool} + \hat{\mu}$$

After Regression:

$$\widehat{GINI} = 46.6 - 1.0 \text{avgschool}$$

N=153 R^2=0.1584

Table 2: Estimation Results-Model 1 Simple Regression				
Variable	**Coefficient (Std. Error)**	**T-value**	**P>|t|**	**H_0: B□ = 0 H_1: B□ ≠ 0**
avgschool	-1.00*** (0.19)	-5.33	0.00	Reject at 1%
constant	46.59*** (1.69)	27.53	0.00	Reject at 1%

*(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)*

See Appendix Output 1 for STATA Output

Model 2: First Multiple Regression:

Equation:

$$\widehat{GINI} = \hat{\beta}_0 + B_1 OECD + \hat{\beta}_2 avgschool + \hat{\beta}_3 medage + \hat{\beta}_4 vulemp + \hat{\beta}_5 loggdppercap + \hat{\beta}_6 govtexpendeduc + \hat{\beta}_7 newhealthexpend + \hat{\beta}_8 FDIinflows + \hat{\mu}$$

After Regression:

$$\widehat{GINI} = 68.6 + 1.89OECD - 0.55avgschool - 0.58medage - 0.12vulemp - 1.99loggdppercap + 0.46govtexpendeduc + 0.03newhealthexpend - 0.009FDIinflows$$

N=117 R^2=0.3952

Table 3: Estimation Results-Model 2 Multiple Regression				
Variable	**Coefficient (Std. Error)**	**T-value**	**P>|t|**	**H_0: B□ = 0 H_1: B□ ≠ 0**
avgschool	-0.55 (0.389)	-1.42	0.157	Fail to Reject at 10%
medage	-0.58*** (0.132)	-4.37	0.000	Reject at 1%
vulemp	-0.12*** (0.044)	-2.68	0.009	Reject at 1%
loggdppercap	-1.99*** (0.705)	-2.83	0.006	Reject at 1%
govtexpendeduc	0.46 (0.427)	1.08	0.284	Fail to Reject at 10%
newhealthexpend	0.03 (0.026)	1.01	0.315	Fail to Reject at 10%
OECD	1.89 (1.797)	1.05	0.295	Fail to Reject at 10%
FDIinflow	-0.009 (0.024)	-0.39	0.694	Fail to Reject at 10%
constant	68.62*** (6.585)	10.42	0.000	Reject at 1%

*(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)*

See Appendix Output 2 for STATA Output

Model 3: Second Multiple Regression:

Equation:

$$\widehat{GINI} = \hat{\beta}_0 + \hat{\beta}_1 \text{ avgschool} + \hat{\beta}_2 \text{ medage} + \hat{\beta}_3 \text{ vulemp} + \hat{\beta}_4 \text{ loggdppercap} + \hat{\mu}$$

After Regression:

$$\widehat{GINI} = 62.27 - 0.23\text{avgschool} - 0.55\text{medage} - 0.08\text{vulemp} - 0.78\text{loggdppercap}$$

N=150 R^2=0.2727

Table 4: Estimation Results-Model 3 Multiple Regression				
Variable	**Coefficient (Std. Error)**	**T-value**	**P>|t|**	**H_0: B□ = 0 H_1: B□ ≠ 0**
avgschool	-0.23 (0.373)	-0.61	0.544	Fail to Reject at 10%
medage	-0.55*** (0.120)	-4.54	0.000	Reject at 1%
vulemp	-0.08** (0.040)	-2.07	0.040	Reject at 5%
loggdppercap	-0.78 (0.621)	-1.26	0.211	Fail to Reject at 10%
constant	62.27*** (5.019)	12.41	0.000	Reject at 1%

*(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)*

See Appendix Output 3 for STATA Output

Model 4: Multiple Regression Model 3:

Equation:

$$\widehat{GINI} = \hat{\beta}_0 + \hat{\beta}_1 \text{ avgschool} + \hat{\beta}_2 \text{ medage} + \hat{\beta}_3 \text{ vulemp} + \hat{\mu}$$

After Regression:

$$\widehat{GINI} = 59.67 - 0.32\text{avgschool} - 0.54\text{medage} - 0.09\text{vulemp}$$

N=150 R^2=0.2648

Table 5: Estimation Results-Model 4 Multiple Regression				
Variable	**Coefficient (Std. Error)**	**T-value**	**P>|t|**	**H_0: B□ = 0 H_1: B□ ≠ 0**
avgschool	-0.32 (0.366)	-0.88	0.383	Fail to Reject 10%
medage	-0.54*** (0.120)	-4.46	0.000	Reject at 1%
vulemp	-0.09** (0.040)	-2.19	0.030	Reject at 5%
constant	59.67*** (4.578)	13.03	0.000	Reject at 1%

*(*Statistically Significant at 10%, **Statistically Significant at 5%, ***Statistically Significant at 1%)*

See Appendix Output 4 for STATA Output

Interpretation:

When beginning the project, we wanted to start with our simple regression model and then build out the models based on omitted variables and significance of the variables that had already been tested. Within each round of regression, we conducted a two-tail test to assess the significance of the variable or variables on the dependent variable as in this case Gini. Once calculated, we omitted insignificant variables as they did not influence our dependent variable as much as we thought they would.

First in Model 1,the Simple Regression model, we tested Average Educational Attainment with regards to the Gini Index. Within this we ran a two-tailed test to assess the significance of Average Educational Attainment on the Gini Index. Once regressed, Average Educational Attainment turned out to be negatively correlated with Gini at -1.00. This draws the conclusion that as Average Educational Attainment rises then the Gini Index decreases so the country becomes closer to equality. This followed

our hypothesis that as a country becomes more educated, income inequality will decrease. Within this regression and the two-tailed test, we also found out that Average Educational Attainment is significant at 1% level. This illustrates that our variable Average Educational Attainment is statistically significant almost all the time. Further, our R^2 was 0.1584 meaning that Average Educational Attainment only explained 15.84% of the Gini Index. With this we decided to add more variables in order to improve this goodness of fit and better explain the Gini Index in order to provide the full story behind income inequality.

Secondly in Model 2, the first Multiple Regression model, we regressed Gini on Average Educational Attainment, Median Age, Government Expenditure on Education, Vulnerable Employment, Health Expenditure, OECD, GDP per Capita, and FDI inflows. We chose numerous variables in order to gain a comprehensive view of income inequality. With this being said, all coefficients for our independent variables were negative except Government Expenditure on Education, Health Expenditure, and OECD. As these increased or took the value of 1 such as in the OECD case, the Gini Index increased, which drew and interesting conclusion because we hypothesized that as a country invests more in itself through education or health then it would become more equal, but the regression does not tell that story. However, when conducting the two tail test we found that these variables in addition to FDI inflows and Average Educational Attainment proved statistically insignificant. Variables such as Median Age, Vulnerable Employment, and GDP per capita proved to be statistically significant at the 1% level. In addition, the R^2 for this model increased to 0.3952. This concludes that the independent variables we listed helped to better explain Gini by almost double.

Thirdly in Model 3, the second Multiple Regression model, we regressed Gini on Average Educational Attainment, Median Age, Vulnerable Employment, and GDP per Capita. We chose these variables because we wanted to remove any insignificant variable from our equation besides are main independent variable of Average Educational Attainment. After running the regression, the coefficients of the variables stayed relatively the same providing the same story as those particular variables increase, a country's Gini index lowers therefore the country becomes more equal. After running the two-tailed test, we concluded that Median Age remained statistically significant at the 1% level and Vulnerable employment remained statistically significant at 5% level. Additionally, our R^2 decreased to 0.2727, which brought up the question of whether the variables we omitted were jointly significant. This eventually led to the F-Test as seen in the Extensions section.

Finally, in Model 4, the third Multiple Regression model, we regressed Gini on Average Educational Attainment, Median Age, and Vulnerable Employment. We removed GDP per capita because it proved to be statistically insignificant in Model 3. Within this regression, we found that the Median Age and Vulnerable Employment remained statistically significant at 5% level. Additionally, our R^2 decreased to 0.2648, which ultimately helped us draw the conclusion that even though the variables we had tested before were not necessarily significant by themselves, they could be jointly significant and therefore help explain the story of income inequality better.

V. Extensions

Robustness Test

Since four of our variables turned out to be statistically insignificant in Model 2, but highly correlated, we decided to run the Robustness Test also known as the F-Test on them in order to test their joint significance. By the F-Test we can determine if these variables have any impact on our model at all and are valuable to the model or not. Our hypothesis is as follows:

$$H_0: \hat{\beta}_1 = \hat{\beta}_4 = \hat{\beta}_5 = \hat{\beta}_7$$

$$H_1: H_0 \text{ is false}$$

Following our hypothesis, we created the restricted model off of Model 2, which served as our unrestricted model.

Unrestricted Model:

$$\widehat{GINI} = \hat{\beta}_0 + \hat{\beta}_1 \text{avgschool} + \hat{\beta}_2 \text{medage} + \hat{\beta}_3 \text{vulemp} + \hat{\beta}_4 \text{loggdppercap} + \hat{\beta}_5 \text{govtexpendeduc} + \hat{\beta}_6 \text{newhealthexpend} + \hat{\beta}_7 OECD + \hat{\beta}_8 \text{FDIinflows} + \hat{\mu}$$

Restricted model:

$$\widehat{GINI} = \hat{\beta}_0 + \hat{\beta}_1 \text{medage} + \hat{\beta}_2 \text{vulemp} + \hat{\beta}_3 \text{newhealthexpend} + \hat{\beta}_4 \text{FDIinflows} + \hat{\mu}$$

We can reject the null hypothesis. Average years of schooling, log of GDP per capita, government expenditure as a percentage of GDP, and membership in the OECD are jointly statistically significant at the 5% level, therefore they are important for our model in explaining income inequality and should be included in the future.

Table 6: F-Test Results	
F(4,108)	Critical Value
11.0004722	2.45

Further, when we first started this project we started out using the variable GDP per capita and did not take the natural log of it. This proved that from the beginning GDP per capita was not statically significant at any level. To combat this insignificance and heteroskedasticity, we took the natural log of GDP per capita to implement a different functional form of the variable to explore if this would help our results. Once the natural log of GDP per capita was taken, it proved statistically significant and helped our regression model significantly therefore demonstrating how the different functional form improved our model.

VI. Conclusions

We began this research project hoping to find out what factors breed income inequality on the country-level; to start, we tested Gini indices on average educational attainment values. This simple linear regression confirms what we expected to see: there is a negative relationship between average years of schooling and income inequality. We expanded and tested out a number of multiple linear regressions in order to capture the multidimensional nature of inequality. The fullest version, Model 2, contains eight independent variables. Only three were statistically significant: the log of GDP per capita, the percentage of vulnerable employment, and the median age. Despite our initial hypothesis, we found that average years of schooling was no longer statistically significant when the other seven variables were held constant.

After taking out 3 variables present in Model 2, our second multiple linear regression model, Model 3, remained relatively the same and yielded similar findings to Model 2. Of the variables in model 3, only two were statistically significant: Median Age and Vulnerable Employment. This model did not advance our findings much further than the previous one. In the final model, Model 4, we conducted a third multiple regression model containing 3 variables: Average Educational Attainment, Median Age and Vulnerable employment. In our final model, we took a closer look at joint significance between variables. Although the variables used in that model were not significant on their own, testing their joint significant helps explain our findings in inequality.

Going forward, we would suggest a more targeted approach to discerning the cause of income inequality. It would be very difficult for a government to take many sweeping actions at once, and it may be more helpful for an econometrician to be able to point out a single, wise, first step that ought to be taken on the road to equality. This could be achieved by first breaking down inequality into dimensions, such as health, education, income level of the country, etc. We attempted this, but our proxies for each dimension were selected largely on the basis of data availability. If a researcher with greater resources and time could develop more appropriate measures for each of these categories, then holding them equal in a regression would be more possible. Afterward, the researcher could select a narrow window of interesting variables that governments have the power to influence, such as quality of education itself. We substituted government expenditure on education as a proxy, but there is likely a better measure if someone were willing to collect the data.

References

Aaberge, Rolf, and Magne Mogstad. “Inequality in Current and Lifetime Income.” *Social Choice and Welfare*, vol. 44, no. 2, 2014, pp. 217–230., doi:10.1007/s00355-014-0838-3.

Abdullah, A. , Doucouliagos, H. and Manning, E. (2015), DOES EDUCATION REDUCE INCOME I INEQUALITY? A META‐REGRESSION ANALYSIS. Journal of Economic Surveys, 29: 301-316. https://onlinelibrary.wiley.com/action/showCitFormats?doi=10.1111%2Fjoes.12056

Annan-Diab, Fatima, and Carolina Molinari. “Interdisciplinarity: Practical Approach to Advancing Education for Sustainability and for the Sustainable Development Goals.” *The International Journal of Management Education*, vol. 15, no. 2, 2017, pp. 73–83., doi:10.1016/j.ijme.2017.03.006.

Cingano, F. (2014), "Trends in Income Inequality and its Impact on Economic Growth", *OECD Social, Employment and Migration Working Papers*, No. 163, OECD Publishing, Paris, https://doi.org/10.1787/5jxrjncwxv6j-en.

Judson, R. (1998, December 4). Economic Growth and Investment in Education: How Allocation Matters. Retrieved October 16, 2019, from https://www.jstor.org/stable/pdf/40215992.pdf?refreqid=excelsior:76df9558db9e8fecce489f9d630f12f3.

Oecd Countries Population. (2019-10-08). Retrieved 2019-10-18, from http://worldpopulationreview.com/countries/oecd-countries/

Park, K. H. (1996). Educational expansion and educational inequality on income distribution. *Economics of Education Review*, *15*(1), 51–58. doi: 10.1016/0272-7757(95)00000-3

Sylwester, K. (2002). A Model of Public Education and Income Inequality with a Subsistence Constraint. *Southern Economic Journal*, *69*(1), 144. doi: 10.2307/1061561

“Table 3: Inequality-Adjusted Human Development Index.” *United Nations Development Programme Human Development Reports*, United Nations Development Programme,

http://hdr.undp.org/en/composite/IHDI.

Vandenbroucke, Guillaume and Zhu, Heting, Aging and Wealth Inequality (2017). Economic Synopses, Issue 2, pp. 1-2, 2017. Available at SSRN: https://ssrn.com/abstract=2925733

Appendix

Table 1-List of Countries included in Dataset(* denotes an OECD member country)

Afghanistan	Central African Republic	Germany*	Lesotho	Palestine, State of	Suriname
Albania	Chad	Ghana	Liberia	Panama	Sweden*
Algeria	Chile*	Greece*	Libya	Papua New Guinea	Switzerland*
Andorra	China	Grenada	Lithuania*	Paraguay	Syrian Arab Republic
Angola	Colombia	Guatemala	Luxembourg*	Peru	Tajikistan
Antigua and Barbuda	Comoros	Guinea	Madagascar	Philippines	Tanzania (United Republic of)
Argentina	Congo	Guinea-Bissau	Malawi	Poland*	Thailand
Armenia	Congo (Democratic Republic of the)	Guyana	Malaysia	Portugal*	The former Yugoslav Republic of Macedonia
Australia*	Costa Rica	Haiti	Maldives	Qatar	Timor-Leste
Austria*	Croatia	Honduras	Mali	Romania	Togo
Azerbaijan	Cuba	Hong Kong, China (SAR)	Malta	Russian Federation	Tonga
Bahamas	Cyprus	Hungary*	Mauritania	Rwanda	Trinidad and Tobago
Bahrain	Czechia*	Iceland*	Mauritius	Saint Kitts and Nevis	Tunisia
Bangladesh	Côte d'Ivoire	India	Mexico*	Saint Lucia	Turkey*

Barbados	Denmark*	Indonesia	Micronesia (Federated States of)	Saint Vincent and the Grenadines	Turkmenistan
Belarus	Djibouti	Iran (Islamic Republic of)	Moldova (Republic of)	Samoa	Tuvalu
Belgium*	Dominica	Iraq	Mongolia	Sao Tome and Principe	Uganda
Belize	Dominican Republic	Ireland*	Montenegro	Saudi Arabia	Ukraine
Benin	Ecuador	Israel*	Morocco	Senegal	United Arab Emirates
Bhutan	Egypt	Italy*	Mozambique	Serbia	United Kingdom*

Bolivia (Plurinational State of)	El Salvador	Jamaica	Myanmar	Seychelles	United States*
Bosnia and Herzegovina	Equatorial Guinea	Japan*	Namibia	Sierra Leone	Uruguay
Botswana	Eritrea	Jordan	Nepal	Singapore	Uzbekistan
Brazil	Estonia*	Kazakhstan	Netherlands*	Slovakia	Vanuatu
Brunei Darussalam	Eswatini (Kingdom of)	Kenya	New Zealand*	Slovenia*	Venezuela (Bolivarian Republic of)
Bulgaria	Ethiopia	Kiribati	Nicaragua	Solomon Islands	Viet Nam
Burkina Faso	Fiji	Korea (Republic of)	Niger	Somalia	Yemen
Burundi	Finland*	Kuwait	Nigeria	South Africa	Zambia
Cabo Verde	France*	Kyrgyzstan	Norway*	South Sudan	Zimbabwe
Cambodia	Gabon	Lao People's Democratic Republic	Oman	Spain*	
Cameroon	Gambia	Latvia*	Pakistan	Sri Lanka	
Canada*	Georgia	Lebanon	Palau	Sudan	

Specific Sources utilized to obtain each variable by the UNDP:

Gini Index: UNDESA (2017a), UNESCO Institute for Statistics (2018), United Nations Statistics Division (2018b), World Bank (2018b), Barro and Lee (2016) and IMF (2018)

Average Educational Attainment: UNDESA (2017a), UNESCO Institute for Statistics (2018), United Nations Statistics Division (2018b), World Bank (2018b), Barro and Lee (2016) and IMF (2018)

Median Age: UNDESA (2017a), UNESCO Institute for Statistics (2018), United Nations Statistics Division (2018b), World Bank (2018b), Barro and Lee (2016) and IMF (2018)

Government Expenditure on Education: World Bank (2018a). World Development Indicators database

Vulnerable Employment: UNDESA (2017a), UNESCO Institute for Statistics (2018), United Nations Statistics Division (2018b), World Bank (2018b), Barro and Lee (2016) and IMF (2018)

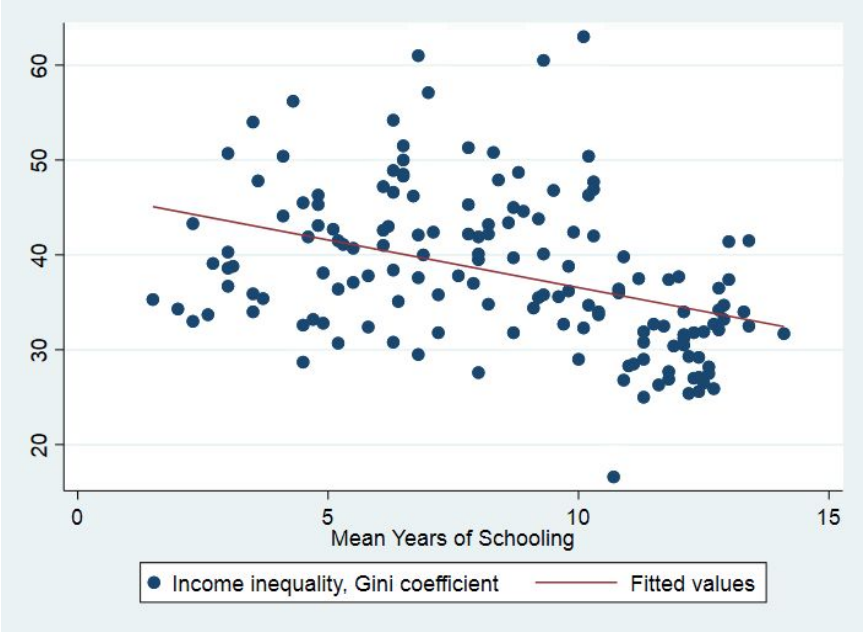
Health Expenditure: UNDESA (2017a), UNESCO Institute for Statistics (2018), United Nations Statistics Division (2018b), World Bank (2018b), Barro and Lee (2016) and IMF (2018)

OECD: https://www.oecd.org/about/members-and-partners/

GDP per Capita: UNDESA (2017a), UNESCO Institute for Statistics (2018), United Nations Statistics Division (2018b), World Bank (2018b), Barro and Lee (2016) and IMF (2018)

FDI inflows: World Bank (2018a). World Development Indicators database

Figure 1-Scatter plot and trend line illustrating Gini and Average Educational attainment regressed



Output 1-Correlation coefficients between independent variables

```
. corr Gini MedAge govtexpendeduc vulemp newhealthexpend OECD avgschool gdppercap FDIinflows
(obs=117)
```

	Gini	MedAge	govtex~c	vulemp	newhea~d	OECD	avgsch~l	gdpper~p	FDIinf~s
Gini	1.0000								
MedAge	-0.5204	1.0000							
govtexpend~c	-0.0320	0.2955	1.0000						
vulemp	0.2988	-0.8082	-0.3892	1.0000					
newhealthe~d	0.1355	0.0093	0.0505	-0.0780	1.0000				
OECD	-0.3060	0.6644	0.2287	-0.5794	-0.0203	1.0000			
avgschool	-0.4485	0.8539	0.3067	-0.8502	0.0162	0.5818	1.0000		
gdppercap	-0.2454	0.1601	0.1910	-0.1894	-0.1699	0.2255	0.2136	1.0000	
FDIinflows	-0.0688	0.0168	-0.0738	0.0581	0.0551	-0.0128	0.0438	-0.0787	1.0000

Output 2-Model 1:Simple Regression Model of Gini vs AvgSchool

```
. regress Gini avgschool
```

Source	SS	df	MS	Number of obs	=	153
				F(1, 151)	=	28.42
Model	1603.21369	1	1603.21369	Prob > F	=	0.0000
Residual	8517.29276	151	56.4059123	R-squared	=	0.1584
				Adj R-squared	=	0.1528
Total	10120.5064	152	66.5822793	Root MSE	=	7.5104

Gini	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
avgschool	-1.003367	.1882029	-5.33	0.000	-1.375218	-.6315155
_cons	46.58708	1.692346	27.53	0.000	43.24334	49.93082

Output 3-Model 2:Multiple Regression Model 1

```
. regress Gini MedAge govtexpendeduc vulemp newhealthexpend OECD avgschool loggd
> ppercap FDIinflows
```

Source	SS	df	MS	Number of obs	=	117
				F(8, 108)	=	8.82
Model	2737.44538	8	342.180672	Prob > F	=	0.0000
Residual	4189.72385	108	38.7937393	R-squared	=	0.3952
				Adj R-squared	=	0.3504
Total	6927.16922	116	59.7169761	Root MSE	=	6.2285

Gini	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MedAge	-.5769628	.1320359	-4.37	0.000	-.8386808	-.3152448
govtexpende~c	.4603302	.4274002	1.08	0.284	-.3868512	1.307512
vulemp	-.1173812	.0438207	-2.68	0.009	-.2042414	-.030521
newhealthex~d	.0258161	.0255597	1.01	0.315	-.0248477	.0764799
OECD	1.890876	1.796963	1.05	0.295	-1.671016	5.452768
avgschool	-.5538529	.3886924	-1.42	0.157	-1.324309	.2166029
loggdppercap	-1.994378	.7051781	-2.83	0.006	-3.392163	-.5965923
FDIinflows	-.0094033	.0238504	-0.39	0.694	-.0566789	.0378723
_cons	68.62002	6.58455	10.42	0.000	55.5683	81.67174

Output 4-Model 3: Multiple Regression Model 2

```
. regress Gini MedAge vulemp avgschool loggdppercap
```

Source	SS	df	MS	Number of obs	=	150
				F(4, 145)	=	13.59
Model	2738.41345	4	684.603362	Prob > F	=	0.0000
Residual	7301.84555	145	50.3575555	R-squared	=	0.2727
				Adj R-squared	=	0.2527
Total	10040.259	149	67.3842886	Root MSE	=	7.0963

Gini	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MedAge	-.5453079	.1201224	-4.54	0.000	-.782725	-.3078908
vulemp	-.0831076	.04017	-2.07	0.040	-.162502	-.0037131
avgschool	-.2269046	.3727946	-0.61	0.544	-.9637181	.5099089
loggdppercap	-.7797516	.620993	-1.26	0.211	-2.007119	.4476159
_cons	62.27485	5.018941	12.41	0.000	52.35512	72.19458

Output 5-Model 4: Multiple Regression Model 3

```
. regress Gini MedAge vulemp avgschool
```

Source	SS	df	MS	Number of obs	=	150
				F(3, 146)	=	17.53
Model	2659.0165	3	886.338834	Prob > F	=	0.0000
Residual	7381.24249	146	50.5564554	R-squared	=	0.2648
				Adj R-squared	=	0.2497
Total	10040.259	149	67.3842886	Root MSE	=	7.1103

Gini	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MedAge	-.5354917	.1201043	-4.46	0.000	-.7728592	-.2981242
vulemp	-.0876488	.0400858	-2.19	0.030	-.1668722	-.0084253
avgschool	-.3205261	.3659831	-0.88	0.383	-1.043835	.4027829
_cons	59.66606	4.577734	13.03	0.000	50.61888	68.71325

Output 6-Restricted Model 2 for F-Tests

```
. regress Gini MedAge vulemp newhealthexpend FDIinflows
```

Source	SS	df	MS	Number of obs	=	149
				F(4, 144)	=	14.72
Model	2910.3637	4	727.590926	Prob > F	=	0.0000
Residual	7116.09355	144	49.4173163	R-squared	=	0.2903
				Adj R-squared	=	0.2706
Total	10026.4573	148	67.7463328	Root MSE	=	7.0297

Gini	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
MedAge	-.6039397	.1012236	-5.97	0.000	-.8040156	-.4038637
vulemp	-.0704557	.0354423	-1.99	0.049	-.14051	-.0004013
newhealthexpend	.0528298	.0260356	2.03	0.044	.0013685	.1042911
FDIinflows	.0077158	.0233528	0.33	0.742	-.0384428	.0538744
_cons	55.26932	4.469334	12.37	0.000	46.43535	64.1033